Detecting change in mean of functional data observations

Nicholas Humphreys

Abstract

We consider a couple of different CUSUM statistics for use in detecting the change in mean of a functional data set. The CUSUM statistics are then used on a real data set that is made up of the daily temperature in England from 1780 through 2007. We want to see whether the mean temperature has changed over the last 228 years. We must first put the data set in a form that can be easily imported into the computer package R.

1 Formatting the data set

The data set used for this project was a very unique one. It consists of the recorded daily temperature in London from 1776-2007. The data set was taken from the website of the British Atmospheric Data Centre. It is one of the most complete recording of temperatures in the world. The data set, although complete, was not in a format that was very easily read into a computer program(R was used in this project). The format of the data looked like this:

year	day	Jan	Feb	⋯	Nov	Dec
1776	1	32	-15	⋯	78	112
1776	2	20	7	⋯	85	62
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1776	31	-8	-999	⋯	-999	22
1777	1	20	0	⋯	68	55
1777	2	10	17	⋯	34	75
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2007	31	63	-999	⋯	-999	65

where the first column is the year, the second column is the day of the month, and the 3rd through 14th columns are the months of the year ranging from January to December. For example, the 3rd row 3rd column is the recorded temperature for January 3rd of

1776. The 5th row 6th column is the temperature for April 5th 1776. Note that these temperatures are listed in tenths degrees celsius. The data set also contains the values -999. These are nothing more than placeholders for the data set. They are input as values for days that don't exist. For example, row 30 column 4 would be February 30th of 1776. It is impossible to have a temperature reading on this day due to the fact that February never has 30 days.

The data matrix above shows that there are $31 \times 14 = 434$ data points for each year with 232 years worth of data. This means there are $434 \times 232 = 100,688$ data points in this data set. R can only handle 99,999 data points in a single matrix, so I had to cut the data set down so that the new data set will only include the years 1780-2007. There were 3 processes that needed to be performed in order to get the data set into something more readable.

First, the data set needed to be put into a new matrix where each row contained the data points for just one year; the resulting matrix was 228 rows(the number of years from 1780-2007) by 372 columns(31 days $\times$ 12 months). This was accomplished by writing a program that made the first 31 data ponts in column 3 of the data set equal to the first 31 days of each row in the newly created matrix. For example, the first 31 data points in the third column became the first 31 data ponts of the first row(January first through 31st of 1780). Data points 32 though 62 of the third column became the first 31 days of row 2 of the new matrix(January 1st through 31st of 1781). This was continued until the new matrix contained the first 31 data points for all 228 rows. The program then moved to the fourth column and repeated the process. For example, the first 31 data points of column 4 became the 32 through 62 data points of row 1 in the new matrix(february 1st through 31st of February). The process continued until all of the data points have been filled in and the resulting matrix is 228 rows by 372 columns. Of course, the reason that there are 372 days in the year is because we have not taken out the -999 yet. The new data set looked like this:

year	day				
	1	2	···	371	372
1780	-26	3	···	89	75
1781	78	22	···	42	53
⋮	⋮	⋮	⋮	⋮	
2006	56	45	···	86	88
2007	71	63	···	60	65

The second problem was leap year. Every four years, starting with 1780, there is a temperature recording for February 29th. For analysis purposes, all of the rows in our data set need to be the same length. This is not possible if every four years there is an extra temperature reading. The best possible solution was to just get rid of all of the values for February 29th. The data value for February 29th was in the 60th column, and as mentioned before occurs every four years. Knowing that I would eventually have to write a program to get rid of the values of -999, I decided the easiest thing to do would be to just change the value of February 29th to -999. The program used for this was very simple. It read in the data matrix and using a for loop assigned the value of -999 to the data value in the 60th column and every fourth row starting with the first row. The new data set looked the same as above except that there is now the value of -999 as the data value for February 29th.

The third problem that needed to be resolved was the existence of the data points with values -999. A program needed to be written to get rid of these data points and shift the data so that the data set just contained the temperature for days that actually existed. This was done by simply writing a program whose input was the 228×372 matrix and whose output was a matrix that only printed values greater than -999; thus eliminating all the values of -999 from the matrix. The new matrix is 228 rows by 365 columns. The new matrix is again the same as above except the values of -999 are deleted. Each row contains the daily temperature from January 1st through December 31st starting with 1780 through 2007. Now that the matrix is in a format that R can read,a few tests can be performed on the data.

2 Detecting changes in the mean

One of the main objectives in this project is to see if the mean temperature in England has changed over the last 228 years. The CUSUM statistic is a method that we can use to detect a change. Let $X_1(t), X_2(t), \ldots, X_n(t)$ for $0 \leq t \leq 1$ be the observation from the data set, that is, each $X_i(t)$ is a random function that give the daily temperature for each year $0 \leq i \leq n$. There is 228 years in the data set so $n = 228$.

We wish to test the null hypothesis

$$H_{0,\mathrm{mean}}: \quad EX_1(t) = EX_2(t) = \cdots = EX_n(t)$$

against the alternative

$$H_{A,\mathrm{mean}}: \text{ there is } 1 < k < n \text{ such that } EX_1(t) = EX_2(t) = \cdots = EX_k(t) \neq EX_{k+1}(t) = \cdots = EX_n(t).$$

The most popular tests for $H_{0,\mathrm{mean}}$ against $H_{A,\mathrm{mean}}$ are based on the CUSUM statistics. The CUSUM process was found in the paper by Berkes, Gabrys, Horváth, and Kokoszka (2007). The CUSUM process is defined as

$$(2.1) \qquad Z_k(j) = \frac{1}{\sqrt{n}\hat{\sigma}_k} \left\{ \sum_{1 \leq i \leq j} c_i(k) \; - \; \frac{j}{n} \sum_{1 \leq i \leq n} c_i(k) \right\} \qquad 1 \leq j \leq n,$$

where $\hat{\sigma_k}^2$ is the sample variance defined by

$$\hat{\sigma_k}^2 = \frac{1}{n-1} \sum_{1 \le i \le j} (c_i(k) \quad - \quad \bar{c}(k))^2 \tag{2.2}$$

and $\bar{c}(k)$ is the average value of the Fourier coefficients for all 228 years defined as

$$\bar{c}_k = \frac{1}{n} \sum_{1 \le i \le n} c_i(k) \tag{2.3}$$

The sequence $\{c_i(k)\}$ is usually referred to as the sequence of Fourier coefficients of $X(t)$ with respect to the orthonormal set $\{\rho_k(t)\}$. $\{c_i(k)\}$ can be constructed for any $\{\rho_k\}$ whether or not they form a basis. This sequence is defined as

$$c_i(k) = \int_0^1 X_i(t)\rho_k(t)dt \quad 1 \le i \le n \tag{2.4}$$

This expression is used to project the random functions with respect to some eigenfunctions $\rho_k(t)$ for $k = 1, \ldots, d$ of Brownian Motion that were derived by solving the expression $\lambda_k \rho_k(t) = \int_0^1 c(t,s)\rho_k(s)ds$ for $c(t,s) = \min(t,s)$. Also, the eigenfunctions need to form an orthonormal system, that is, they need to satisfy the conditions

(i) $\int_0^1 \rho_k^2(t)\, dt = 1$.

(ii) $\int_0^1 \rho_k(t)\rho_\ell(t)\, dt = 0$ where $k \neq \ell$

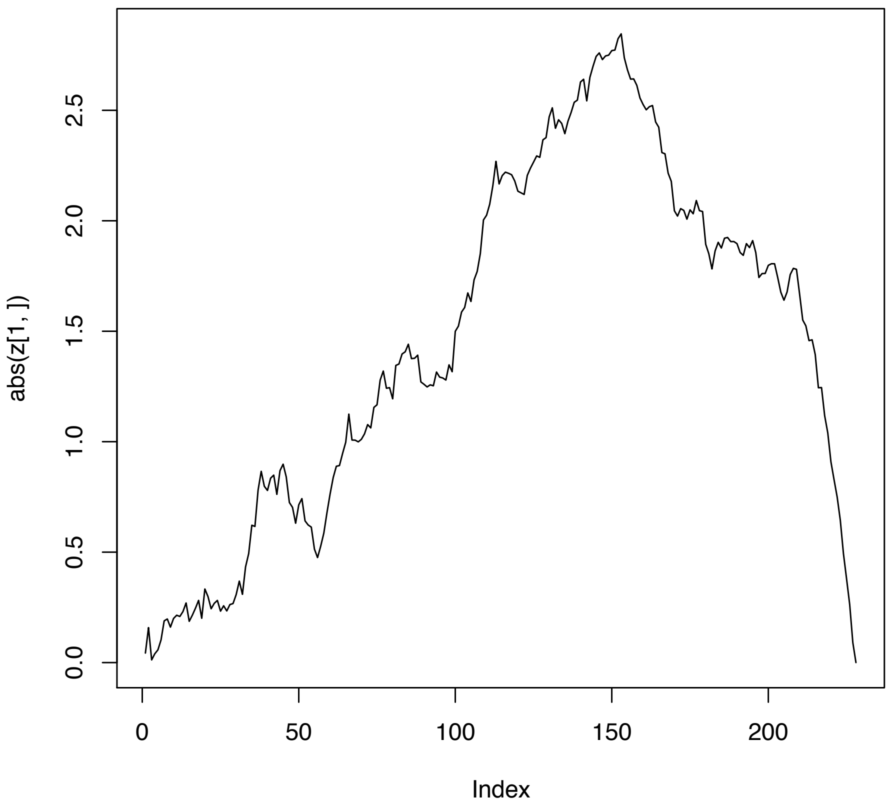
This was solved in the research project from last semester an we found that the eigenfunctions were of the form

(i) $\rho_k(t) = \sqrt{2}\cos(k\pi t)$ for all $k = 1, 2, \ldots$

and

(ii) $\rho_k(t) = \sqrt{2}\sin(k\pi t)$ for all $k = 1, 2, \ldots$

The eigenfunction we chose to use was $\rho_k(t) = \sqrt{2}\sin(k\pi t)$ for $k = 1, 2, 3$. Now that we have everything we need, we can plug everything into the expression for $Z_k(j)$ and get a set of 228 data points for each of the three eigenfunctions. The plot of $|Z_1(j)|$ is shown below.



The critical values were found in Bain and Engelhardt(1992). They are 1.224 = 90%, 1.36 = 95%, and 1.63 = 99%. This means that 1% of our data can be above 1.63 and we would still accept the null hypothesis, otherwise we reject. Looking at the graph of $Z_1(j)$ we can see that a large enough number of the data points lie above the critical value, 46.9% of the data points in fact, that we can reject the null hypothesis that the average temperature in England has remained the same over the last 228 years.

There is another way that we can check the null hypothesis. The first CUSUM expression we used just took into account the projection of one eigenfunction. We now want to combine the three projections into one. This is by the use of another CUSUM expression

$$Q(j) = \frac{1}{n}\Big(\sum_{1\le i\le j}(\underline{A}_i - \overline{\underline{A}})\Big)^T \underline{\hat{D}}^{-1}\Big(\sum_{1\le i\le j}(\underline{A}_i - \overline{\underline{A}})\Big) \quad 1 \le j \le n. \tag{2.5}$$

There are a few expressions in here that need defining. The first of which is $\underline{A}_i$. This is a vetor of the projections of $X_i(t)$ into the linear space of $\{\rho_1(t), \rho_2(t), \rho_3(t)\}$, that is, it is a matrix that contains all of the fourier coefficients, $c_i(k)$, for $1 \le i \le 228$ and $1 \le k \le 3$. It looks like this

$$\begin{bmatrix} c_1(1) & c_2(1) & c_3(1) & \cdots & c_{228}(1) \\ c_1(2) & c_2(2) & c_3(2) & \cdots & c_{228}(2) \\ c_1(3) & c_2(3) & c_3(3) & \cdots & c_{228}(3) \end{bmatrix}$$

$\underline{\hat{D}}$ is the empirical covariance matrix, which measures the dependency between the random variables, is defined as

$$\underline{\hat{D}}(i,j) = \frac{1}{228} \sum_{1 \le \ell \le 228} (c_\ell(i) - \overline{c}(i))(c_\ell(j) - \overline{c}(j)) \quad 1 \le i,j \le 3 \tag{2.6}$$

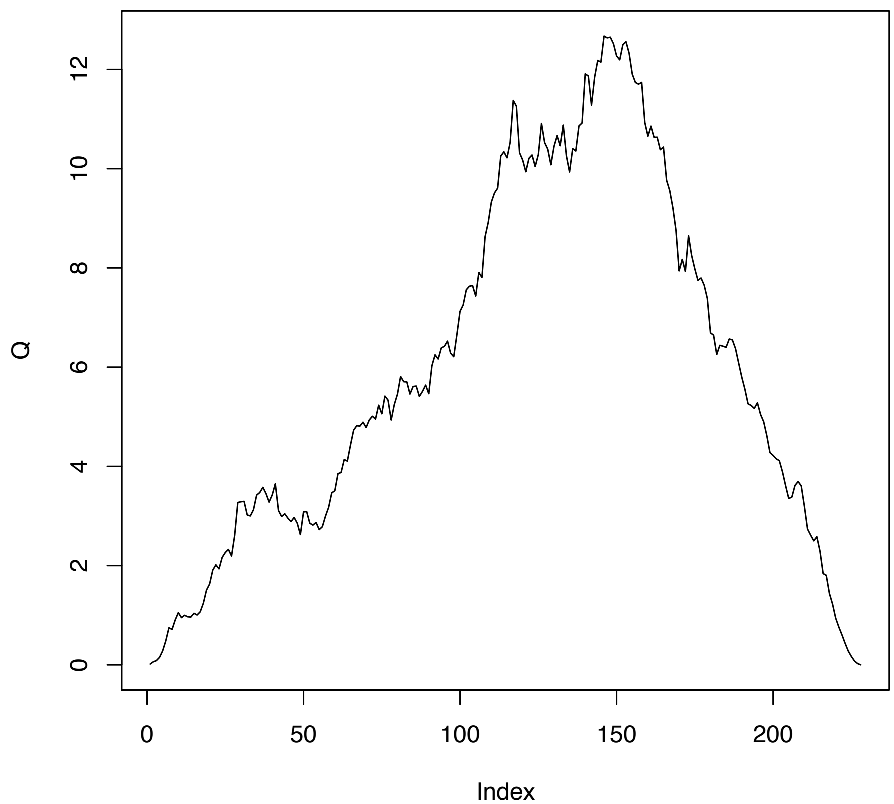
In our case, the result is a 3 by 3 matrix.
Finally, $\underline{\overline{A}}$ is just the vector of the average of all the Fourier coefficients for each projection, that is,

$$\overline{c}(k) = \frac{1}{228} \sum_{1 \le i \le 228} c_i(k) \quad k = 1, 2, 3. \tag{2.7}$$

we write $\underline{\overline{A}}$ as

$$\begin{bmatrix} \overline{c}(1) \\ \overline{c}(2) \\ \overline{c}(3) \end{bmatrix}$$

Now that we have all of the expressions we need, we can use R to evaluate the CUSUM expression Q(j) as defined earlier. The plot of Q(j) is displayed below.



Assuming the same critical values as before, we would also definitely reject the null hypothesis here too.

3 Results and future plans

Both of the CUSUM statistics confirm the alternative hypothesis that the average temperature in England over the last 228 years has not stayed the same. What has caused the change cannot be determined by the given data. In the news media today we hear a lot about global warming as the cause for the change in temperature. Whether this is the cause or not, we would also need to look at what was going on in the England area around the times of the greatest change in temperature. If we look at the time of the Industrial Revolution, we would certainly see that the temperature is increasing, not do to gloabl warming, but due to smoke and heat being released into the atmosphere by the machines invented during that time.

If I have the opportunity to continue this project I would like to use a CUSUM expression that detects the change in variance. I would choose the null hypothesis that the variance from one year to the next has stayed the same over time; and compare it to the alternative hypothesis that the variance is different for at least one year of the data set. I believe that more information can be gathered from this statistic. The reason I believe this is because I feel the CUSUM for change in mean can be somewhat misleading. If we look at the temperature data, we see that on the same day over the 228 year period the temperatures go throug extreme highs and lows, but if we take the average of those temperatures its still somewhere in the middle. I think it would be interesting to see how the variance has changed, if any, over the last 228 years. I personally feel that result will be that the variance is increasing as time goes on, that is, the range of the temperatures was relatively small in the beginning but as time went on the range of temperatures for the same day over different years is getting bigger. The CUSUM statistic that I would use to test change in variance would be one found in the third page of the changepoint paper by Berkes, Gombay and Horváth (2007). This process is defined as

$$M_n^{(r)}(t) = \begin{cases} n^{-1/2}\Bigg(\sum_{1 \le i \le (n+1)t} (X_i - \overline{X}_n)(X_{i-r} - \overline{X}_n) \\ \qquad\qquad -t \sum_{1 \le i \le n} (X_i - \overline{X}_n)(X_{i-r} - \overline{X}_n) \Bigg), & \text{if } 0 \le t < 1 \\ 0, & \text{if } t = 1, \end{cases}$$

where $\overline{X}_n = (1/n)\sum_{1 \le i \le n} X_i$.

I believe that critical values for the data can be found so we can use the same approach as before to accept or reject the null hypothesis that the variance in temperature has been changing over time.

References

[1] Bain, L. J. and Engelhardt, M. (1992). *Introduction to Probability and Mathematical Statistics* (2nd ed.), Duxbury, California.

[2] Berkes, I., Gabrys, R., Horváth, L., Kokoszka, P., (2007). *Detecting changes in mean of functional observations.* Preprint.

[3] Berkes, I., Gombay, E., Horváth, L., (2007). *Testing for changes in the covariance structure of linear processes.* Preprint.